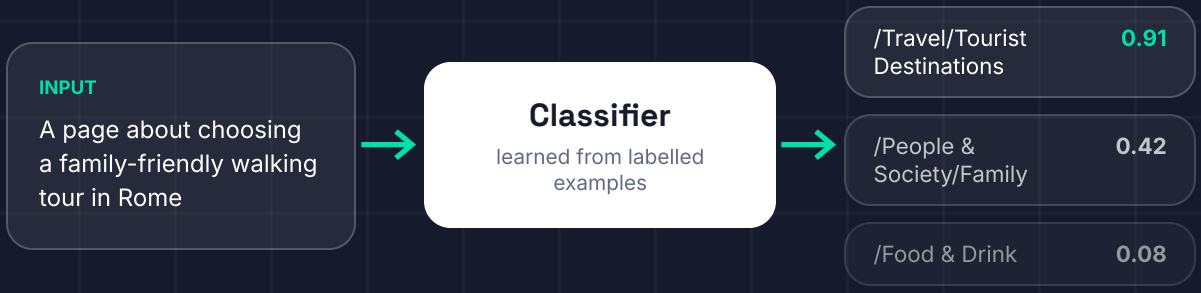
What Is a Classifier?

A plain-language lesson on how machine-learning classifiers sort content, why search systems use them, and what that means for how you write and structure pages.

Teaching series · Lesson 1 · beginner to intermediate

A CLASSIFIER IN ONE PICTURE



Illustrative output. Labels and scores are examples, not a real API response.

40	6	1
minutes	key terms	hands-on exercise

Ameneh Saeednia
Senior SEO and Generative Search Strategist

Before you start

What this lesson covers

Search systems don't read pages the way people do. Much of what they decide, **what a page is about, whether it's spam, which category it belongs in**, comes from classifiers. You don't need to code to understand them.

By the end of this lesson you can

- Explain **machine learning**, **supervised learning** and **vectors** in one sentence each.
- Describe what a **classifier** does, and the difference between **binary, multiclass and multilabel** classification.
- Name where search engines have **publicly said** they use classifier-style systems.
- Run a **simple classification test** on your own page and interpret the result carefully.

Who it's for

SEOs, content writers and marketers who hear “classifier” in Google announcements and want a **working mental model**, not a math course.

How to use it

Read pages 3 to 6, do the **exercise on page 7**, then check yourself with the **quiz on page 8**. Answers are on the last page.

Part 1

Three ideas you need first

Classifiers sit on top of three ideas. Get these and the rest follows.

1

Machine learning

A program **learns** when its performance on a task improves with experience. Tom Mitchell's classic definition: experience E, task T, performance measure P. [1]

2

Supervised learning

The program learns from **labelled examples**: inputs paired with the right answer. Here are 10,000 emails marked spam or not spam; now predict the next one.

3

Vectors

Before a model can learn from text, the text is turned into **a list of numbers**. Similar meanings end up as similar lists. That's what “embedding” means.

Teaching tip

A vector in physics has **direction and size**, like a car heading northwest at 65 mph. In machine learning, think of it as **coordinates**: each page becomes a point in space, and pages about similar things sit close together.

Part 2

What a classifier does

A classifier **assigns input to one or more classes**. It learns the pattern from labelled training data, then applies it to new data it has never seen.

Type	Classes	Example
Binary	Two	Spam or not spam
Multiclass	Several, pick one	A news article is sports, politics or technology
Multilabel	Several, pick any	A film is comedy **and** drama **and** action

1 **Prepare**
Clean the data, turn it into vectors

2 **Select**
Choose the most useful features

3 **Train**
Adjust the model to reduce errors

4 **Evaluate**
Accuracy, precision, recall

5 **Predict**
Classify new, unseen input

Common classifier families

Decision trees, random forests, Naive Bayes, logistic regression, support vector machines, k-nearest neighbours and **neural networks**. Modern text classification mostly uses neural networks.

Where they go wrong

Overfitting (memorizing the training data), **imbalanced data** (too few examples of one class), **noisy labels** and **scale**. A classifier is only as good as its examples.

Part 3

Where search uses classifiers

Search engines don't publish their models, but they have **described** several systems in classifier terms. Stick to what's been said publicly.

System (as described by Google)	What it classifies	What it means for you
Helpful content system (2022 launch)	Google said the process was **automated, using a machine-learning model**, and produced a site-wide signal. [2]	Patterns across a site matter, not just one page.
SpamBrain	Google calls it its **AI-based spam-prevention system**. [3]	Scaled, low-value patterns are detectable as patterns.
Natural Language API: content classification	Google Cloud's public API **assigns content categories** to text with confidence scores. [4]	A useful proxy test. **Not** Search's own classifier.

A careful way to say it

“Google uses machine-learning classifiers for some quality and spam decisions” is supported. “Google's classifier gives my page a score of 0.73” is **not**. Nobody outside Google sees Search's internal scores.

Part 4

What you can actually influence

You can't see Search's classifiers. You can make your page **easy to classify correctly**.

1

Be clear about the subject

Name the main **entity** early and consistently. Ambiguous pages produce low-confidence or wrong categories.

2

Keep one job per page

A page that mixes three topics looks like three weak classes. **Focus** gives a clearer signal.

3

State facts as relationships

“[Tour] is a 3-hour walking tour of [district] for families” gives a system a **subject, attribute and relationship** to work with.

4

Watch the site-wide pattern

If classifiers can work at site level, **thin or off-topic sections** affect more than their own URLs.

Common misconception

Adding more keywords doesn't “convince” a classifier. Models learn **patterns of meaning**, so repetition adds little and can read as low quality.

Exercise

Classify your own page

Time: 15 minutes. You'll use a public demo of a text classifier as a **proxy**, not as a view of Google Search.

Step	What to do
1	Pick one of your pages. Copy its **main body text** (not menus or footers).
2	Paste it into the **Google Cloud Natural Language** demo and run content classification. [4]
3	Write down the **top categories and confidence scores**.
4	Ask: is that the category **you intended**? Is the top score clearly ahead of the second?
5	Rewrite the **first two paragraphs** to name the subject and its purpose more plainly. Run it again.
6	Record both results. Note **what changed and what didn't**.

Interpreting it honestly

A higher confidence score means **this API** reads the text more clearly. It doesn't mean Google ranks the page higher. Treat it as a **clarity check**.

Discussion question

If the API put your page in the wrong category, what in the text could have caused it? **Ambiguous names? Mixed topics? Missing context?**

Check your understanding

Quick quiz

Answers are on the next page.

1. A model that sorts emails into spam or not spam is a:

a) Multilabel classifier
b) Binary classifier
c) Regression model

2. Supervised learning needs:

a) Labelled training examples
b) No data at all
c) Only unlabelled text

3. Which statement about Google's helpful content system is supported by Google's own description?

a) It gave each page a public score
b) It was automated, using a machine-learning model
c) It only checked word count

4. Your page gets a 0.91 “Travel” score in the Natural Language API. This means:

a) It will rank for travel queries
b) The API reads it clearly as travel content
c) Google Search uses the same score

5. Which change is most likely to help a page be classified correctly?

a) Repeating the main keyword more often
b) Naming the main subject early and keeping one topic per page
c) Adding more categories to the footer

Answers and sources

Answer key, glossary and further reading

Answer key

1: **b**. Two classes.
2: **a**. Labelled examples.
3: **b**. Google's own description.
4: **b**. It's a property of that API, not of Search.
5: **b**. Clarity and focus, not repetition.

Glossary

Class: a category a model can assign.
Feature: a measurable property of the input.
Training data: labelled examples a model learns from.
Overfitting: learning the examples instead of the pattern.
Confidence score: how sure a model is about one class.
Embedding: text turned into a vector.

01 **Machine Learning. Tom M. Mitchell, McGraw Hill, 1997**
Source of the experience, task and performance definition.

02 **What creators should know about Google's August 2022 helpful content update. Google Search Central Blog**
Describes the system's automated, machine-learning-based classifier process.
https://developers.google.com/search/blog/2022/08/helpful-content-update

03 **Google Search Central: spam updates and SpamBrain**
Google's description of SpamBrain as an AI-based spam-prevention system.
https://developers.google.com/search/updates/spam-updates

04 **Content classification. Google Cloud Natural Language API**
Public API used in the exercise. Separate from Google Search.
https://cloud.google.com/natural-language/docs/classifying-text

Lesson by Ameneh Saeednia, adapted from 2023 training notes in my archive and prepared with AI assistance. Related reading on seogirl.ca: *Testing content with the Google Natural Language API*.